Exhibit 17

JAMA Guide to Statistics and Methods

Multiple Imputation
A Flexible Tool for Handling Missing Data

Peng Li, PhD; Elizabeth A. Stuart, PhD; David B. Allison, PhD

In this issue of *JAMA*, Asch et al[1] report results of a cluster randomized clinical trial designed to evaluate the effects of physician financial incentives, patient incentives, or shared physician and patient incentives on low-density lipoprotein cholesterol (LDL-C) levels among patients with high cardiovascular risk. Because 1 or more follow-up LDL-C measurements were missing for approximately 7% of participants, Asch et al used multiple imputation (MI) to analyze their data and concluded that shared financial incentives for physicians and patients, but not incentives to physicians or patients alone, resulted in the patients having lower LDL-C levels. Imputation is the process of replacing missing data with 1 or more specific values, to allow statistical analysis that includes all participants and not just those who do not have any missing data.

Related article page 1926

Missing data are common in research. In a previous JAMA Guide to Statistics and Methods, Newgard and Lewis[2] reviewed the causes of missing data. These are divided into 3 classes: (1) missing completely at random, the most restrictive assumption, indicating that whether a data point is missing is completely unrelated to observed and unobserved data; (2) missing at random, a more realistic assumption than missing completely at random, indicating whether a missing data point can be explained by the observed data; or (3) missing not at random, meaning that the missingness is dependent on the unobserved values. Common statistical methods used for handling missing values were reviewed.[2] When missing data occur, it is important to not exclude cases with missing information (analyses after such exclusion are known as complete case analyses). Single-value imputation methods are those that estimate what each missing value might have been and replace it with a single value in the data set. Single-value imputation methods include mean imputation, last observation carried forward, and random imputation. These approaches can yield biased results and are suboptimal. Multiple imputation better handles missing data by estimating and replacing missing values many times.

Use of Method

Why Is Multiple Imputation Used?

Multiple imputation fills in missing values by generating plausible numbers derived from distributions of and relationships among observed variables in the data set.[3] Multiple imputation differs from single imputation methods because missing data are filled in many times, with many different plausible values estimated for each missing value. Using multiple plausible values provides a quantification of the uncertainty in estimating what the missing values might be, avoiding creating false precision (as can happen with single imputation). Multiple imputation provides accurate estimates of quantities or associations of interest, such as treatment effects in randomized trials, sample means of specific variables, correlations between 2 variables, as well as the related variances. In doing so, it reduces the chance of false-positive or false-negative conclusions.

Multiple imputation entails 2 stages: (1) generating replacement values ("imputations") for missing data and repeating this procedure many times, resulting in many data sets with replaced missing information, and (2) analyzing the many imputed data sets and combining the results. In stage 1, MI imputes the missing entries based on statistical characteristics of the data, for example, the associations among and distributions of variables in the data set. After the imputed data sets are obtained, in stage 2, any analysis can be conducted within each of the imputed data sets as if there were no missing data. That is, each of the "filled-in" complete data sets is simply analyzed with any method that would be valid and appropriate for addressing a scientific question in a data set that had no missing data.

After the intended statistical analysis (regression, *t* test, etc) is run separately on each imputed data set (stage 2), the estimates of interest (eg, the mean difference in outcome between a treatment and a control group) from all the imputed data sets are combined into a single estimate using standard combining rules.[3] For example, in the study by Asch et al,[1] the reported treatment effect is the average of the treatment effects estimated from each of the imputed data sets. The total variance or uncertainty of the treatment effect is obtained, in part, by seeing how much the estimate varies from one imputed data set to the next, with greater variability across the imputed data sets indicating greater uncertainty due to missing data. This imputed-data-set-to-imputed-data-set variability is built into a formula that provides accurate standard errors and, thereby, confidence intervals and significance tests for the quantities of interest, while allowing for the uncertainty due to the missing data. This distinguishes MI from single imputation.

Combining most parameter estimates, such as regression coefficients, is straightforward,[4] and modern software (including R, SAS, Stata, and others) can do the combining automatically. There are some caveats as to which variables must be included in the statistical model in the imputation stage, which are discussed extensively elsewhere.[5]

Another advantage of adding MI to the statistical toolbox is that it can handle interesting problems not conventionally thought of as missing data problems. Multiple imputation can correct for measurement error by treating the unobserved true scores (eg, someone's exact degree of ancestry from a particular population when there are only imperfect estimates for each person) as missing,[6] generate data appropriate for public release while ensuring confidentiality,[7] or make large-scale sampling more efficient through planned missing data (ie, by intentionally measuring some variables on only a subset of participants in a study to save money).[8]

What Are the Limitations of Multiple Imputation?

As with any statistical technique, the validity of MI depends on the validity of its assumptions. But when those assumptions are met, MI rests on well-established theory.[3,5] Moreover, substantial empirical support exists for the validity of MI in simulations, including those based on real data patterns.[9] In principle, computational speed can be a problem because each analysis must be run multiple times, but in practice, this is rarely an issue with modern computers.

Many nonstatisticians chafe at "making up data" as is done in MI and note that the validity of MI depends on an assumption about which factors relate to the probability that a data point is missing. Because of concern this assumption may be violated, it is tempting to retreat to the safe haven of complete case analysis, ie, only analyze the participants without missing values. This safe haven is, however, illusory. Although rarely made explicit by users, complete case analysis requires a far more restrictive assumption: that any data point missing is missing completely at random. Other common strategies—mean imputation, last observation carried forward, and other single imputation approaches—underestimate standard errors by ignoring or underestimating the inherent uncertainty created by missing data, a problem MI helps overcome.

Why Did the Authors Use Multiple Imputation in This Particular Study?

In the study by Asch et al,[1] the primary outcome, LDL-C levels, had missing values. Thus, a method to handle missingness was needed to maintain the validity of the statistical inferences. Complete case analysis would have inappropriately not included 7% of their sample, leading to less study power, results restricted to those individuals without missing values, violation of the intent-to-treat principle, possible nonrandom loss and therefore a loss of the ability to rely on the fact of randomization to justify causal inferences, and ultimately to results that may not apply to the original full sample.

How Should Multiple Imputation Findings Be Interpreted in This Particular Study?

Provided that the underlying assumptions of MI are met, the results from this study can be interpreted as if all the participants had no missing entries. That is, both the estimates of quantities like means and measures of association and the estimates of their uncertainty (standard errors) on which formal statistical testing is based will not be biased by the fact that some data were missing. There would have been greater precision of the estimates and study power had there been no missing data. But imputation at least appropriately reflects the amount of information there actually is in the data available.

Caveats to Consider When Looking at Results Based on Multiple Imputation

When the missing data are not missing at random, results from MI may not be reliable. Generally, reasons for missingness cannot be fully identified. In practice, collecting more information about study participants may help identify why data are missing. These "auxiliary variables" can then be used in the imputation process and improve MI's performance. All other things being equal, imputation models with more variables included and a large number of imputations improve MI's performance. Multiple imputation is arguably the most flexible valid missing data approach among those that are commonly used.

ARTICLE INFORMATION

Author Affiliations: Office of Energetics and Nutrition Obesity Research Center, University of Alabama at Birmingham (Li, Allison); Department of Biostatistics, School of Public Health, University of Alabama at Birmingham (Li, Allison); Departments of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Stuart).

Corresponding Author: David B. Allison, PhD, Office of Energetics and Nutrition Obesity Research Center, University of Alabama at Birmingham, Ryals Bldg, Room 140J, 1665 University Blvd, Birmingham, AL 35294 (dallison@uab.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

Funding/Support: This work was supported in part by grants from the National Institutes of Health (NIH) (R25HL124208, R25DK099080, R01MH099010, and P30DK056336).

Role of the Funder/Sponsor: The funding sources had no role in the preparation, review, or approval of the manuscript.

Disclaimer: The opinions expressed are those of the authors and do not necessarily represent those of the NIH or any other organization.

REFERENCES

1. Asch DA, Troxel AB, Stewart WF, et al. Effect of financial incentives to physicians, patients, or both on lipid levels: a randomized clinical trial. *JAMA*. doi:10.1001/jama.2015.14850.

2. Newgard CD, Lewis RJ. Missing data: how to best account for what is not known. *JAMA*. 2015;314(9):940-941.

3. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley; 1987.

4. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.

5. Schafer JL. *Analysis of Incomplete Multivariate Data*. New York, NY: Chapman & Hall; 1997.

6. Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. *Hum Hered*. 2009;68(1):65-72.

7. Wang H, Reiter JP. Multiple imputation for sharing precise geographies in public use data. *Ann Appl Stat*. 2012;6(1):229-252.

8. Capers PL, Brown AW, Dawson JA, Allison DB. Double sampling with multiple imputation to answer large sample meta-research questions: introduction and illustration by evaluating adherence to two simple CONSORT guidelines. *Front Nutr*. 2015;2:6.

9. Elobeid MA, Padilla MA, McVie T, et al. Missing data in randomized clinical trials for weight loss: scope of the problem, state of the field, and performance of statistical methods. *PLoS One*. 2009;4(8):e6624.